

Whole-brain substitute CT generation using Markov random field mixture models

Anders Hildeman¹, David Bolin¹, Jonas Wallin², Adam Johansson³, Tufve Nyholm⁴, Thomas Asklund⁴, and Jun Yu⁵

¹Department of mathematical sciences, Chalmers University of Technology and University of Gothenburg, Sweden

²Department of Statistics, Lund University, Sweden

³Department of Radiation Oncology, University of Michigan, USA

⁴Department of Radiation Sciences, Umeå University, Sweden

⁵Department of Mathematics and Mathematical Statistics, Umeå University, Sweden

September 29, 2016

Abstract

Computed tomography (CT) equivalent information is needed for attenuation correction in PET imaging and for dose planning in radiotherapy. Prior work has shown that Gaussian mixture models can be used to generate a substitute CT (s-CT) image from a specific set of MRI modalities. This work introduces a more flexible class of mixture models for s-CT generation, that incorporates spatial dependency in the data through a Markov random field prior on the latent field of class memberships associated with a mixture model. Furthermore, the mixture distributions are extended from Gaussian to normal inverse Gaussian (NIG), allowing heavier tails and skewness. The amount of data needed to train a model for s-CT generation is of the order of 10^8 voxels. The computational efficiency of the parameter estimation and prediction methods are hence paramount, especially when spatial dependency is included in the models. A stochastic Expectation Maximization (EM) gradient algorithm is proposed in order to tackle this challenge. The advantages of the spatial model and NIG distributions are evaluated with a cross-validation study based on data from 14 patients. The study show that the proposed model enhances the predictive quality of the s-CT images by reducing the mean absolute error with 17.9%. Also, the distribution of CT values conditioned on the MR images are better explained by the proposed model as evaluated using continuous ranked probability scores.

1 Introduction

Ionizing radiation undergo attenuation as it passes through organic tissue. That attenuation affects the dose deposition in radiotherapy and the image acquisition in positron emission tomography (PET). In both cases, the attenuation has to be estimated. The simulation of the dose distribution in radiotherapy makes it possible to optimize the treatment for the individual patient, maximizing the dose to the tumor while keeping the dose to healthy surrounding tissue within acceptable limits. In PET, knowledge of the attenuation in the patient is a prerequisite for accurate quantification of the tracer uptake.

Computed Tomography (CT) X-ray imaging uses the attenuation of X-rays in order to construct a three dimensional image of the interior of the region of interest. Therefore, patients usually undergo a CT scan before radiotherapy treatment or in connection to the PET scan in order to acquire information about the attenuation. It has been shown that it is possible to derive similar attenuation information by the use of magnetic resonance imaging (MRI) [16, 26]. Acquiring such a substitute CT (s-CT) image without exposing the patient to X-ray radiation has some advantages compared to performing a CT scan. Firstly, MRI does not expose the subject to ionizing radiation, which has an inherent risk of damaging tissue. Secondly, MRI information is often of interest for other reasons, for instance in order to increase the soft-tissue contrast and perform motion correction in PET images [5].

MR images do not map to CT images directly so in order to generate a s-CT image some prediction method is needed. There are two categories of methods, Atlas based and machine-learning based [5]. The Atlas based methods finds a geometrical mapping between the MR images of the subject and those of MR images in a template library using image registration techniques. From the templates, CT images are then inversely mapped back to the subject and fused to give a s-CT image. The machine-learning based methods instead learn a mapping between the intensity values in the MR and CT images. This mapping is learned on training data where both MR and CT information is available. The mapping can then be applied for construction of s-CT images based solely on some corresponding MR images.

Johansson et al. [16] took the machine-learning based approach to this problem and utilized Gaussian mixture models (GMM) to map between MR and CT images. The parameters of the model were estimated using an Expectation-Maximization (EM) algorithm [9], and the s-CT images were constructed using the expected value of the CT field conditioned on the available MR images. This method has the advantage that it is not dependent on an image registration step which could compromise prediction results. The model is also quite general and easy to estimate and has been shown to provide accurate results both for dose calculation in radiotherapy and attenuation correction in PET imaging [20, 18].

The voxel values of the CT- and MR-images does not only depend on each other pointwise, there are also spatial dependencies that should be taken into account in a statistical model. Johansson et al. [17] added a spatial component to the GMM approach by incorporating the spatial coordinates of the voxels as auxiliary dimensions of the data. Through this, each mixture class was given a spatial location. The spatial model showed improvements in the post-nasal cavities and inner ear where there is air and bone tissue

in close proximity to each other. However, a problem with giving mixture classes a spatial location is that areas separated in space but of the same tissue type needs to be modeled by different classes. This might yield problems with overfitting and unstable estimates due to the increasing demand for training data. Furthermore, the model does not make use of any spatial interaction between voxels and the mapping of the coordinates has the same drawbacks as Atlas based techniques, i.e. sensitivity to misregistration and abnormal anatomies.

In this paper we take another approach to modeling the spatial dependency. A spatial Markov random field (MRF) model [30, Chapter 4] is applied as a prior distribution for the latent field of class memberships. This will bias voxel classification towards spatial clustering in a way that conveys local spatial structures without putting any global restrictions on the spatial location of the class distributions. Furthermore, the mixture model is extended by using a multivariate normal inverse Gaussian distribution (NIG) [2] for the class distributions. The NIG distribution adds flexibility since it allows for skewness and variable kurtosis which might reduce the number of classes needed to model heavy tailed or skewed data.

A problem with these more flexible models is that they are computationally more demanding to estimate than the GMM. Using models that incorporates spatial dependency for large datasets is not easy, and this is commonly referred to as the “Big N” problem in spatial statistics. In spatial statistics, datasets are usually considered to be big if they contain more than 10^4 measurements. In this application, each image consists of more than 7 million voxels and during the learning phase five such images are needed per patient. Furthermore, a number of patients should be used in order to acquire reliable prediction parameters and all voxels need to be processed in each iteration of the learning algorithm. Thus, computational efficiency of the proposed methods is paramount. Here we introduce a novel approach utilizing the EM gradient algorithm [19] and Gibbs sampling to successfully and efficiently estimate parameters for mixture models including spatial dependency even in data rich environments such as this.

The remainder of this paper is divided in to four sections. Section 2 describes the MRF models and Section 3 introduces the method proposed in order to estimate those models. Section 4 describes a cross-validation study based on data from 14 patients, where the proposed models are compared with the original GMM. Results show that the new spatial model increases the predictive quality in comparison with the original model. Finally, Section 5 presents the conclusions and a discussion of future work. There is also an appendix giving some further details and derivations.

2 Statistical modeling of CT-MR interdependence

In order to find and make use of the dependency between a CT image and the corresponding MR images we will assume a parametric model. Since the data we consider consists of three-dimensional bit-mapped digital images one can consider each voxel (three-dimensional pixel) as a point on a three-dimensional equally spaced lattice. Let us enumerate these voxels $i \in \{1, \dots, N\}$. We model the voxel values $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ as a random field on this discrete grid. Furthermore, since the data consists of one CT images and four MR images, each voxel value is five-dimensional, i.e. each \mathbf{X}_i is five-dimensional.

2.1 Mixture model

This paper extends the work of Johansson et al. [16] which used a GMM in order to model the interdependence between the MR images and the CT image. The probability density function of a general mixture model on \mathbb{R}^d is $f(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i)\pi_k$, where f_k is the density function of the distribution associated to class k , π_k is the prior probability that \mathbf{X}_i belongs to class $k \in \{1, \dots, K\}$ where K is the number of classes. A GMM is obtained if the f_k are chosen as Gaussian, which is the most common choice in the literature.

Let us denote the CT value of voxel i as X_i^A and a vector of the four MR values for voxel i as \mathbf{X}_i^B . We will model the voxel values for all five images jointly as $\mathbf{X}_i = [X_i^A, \mathbf{X}_i^B]$ with a mixture model. Furthermore, we denote the set of the whole random field instead of just a single voxel by omitting the voxel index i , such as $\{\mathbf{X}_i\}_i = \mathbf{X} = [\mathbf{X}^A, \mathbf{X}^B]$.

Constructing a s-CT image from the MR images is equivalent to acquiring a prediction of the CT image from a realization of the random field \mathbf{X}^B . One predictor that we will use is the conditional expectation $\mathbb{E}[X_i^A|\mathbf{X}^B]$, which for a mixture model is

$$\mathbb{E}[X_i^A|\mathbf{X}^B] = \sum_{k=1}^K \mathbb{E}[X_i^A|\mathbf{X}^B, Z_i = k] \mathbb{P}(Z_i = k|\mathbf{X}^B). \quad (1)$$

Here $Z_i \in \{1, \dots, K\}$ is a latent variable that describes which mixture class voxel i belongs to. As a measure of uncertainty of the prediction, we will use the conditional covariance,

$$\begin{aligned} \mathbb{C}[X_i^A|\mathbf{X}^B] &= \mathbb{E}\left[X_i^A (X_i^A)^T \middle| \mathbf{X}^B\right] - \mathbb{E}[X_i^A|\mathbf{X}^B] \mathbb{E}[X_i^A|\mathbf{X}^B]^T \\ &= \sum_{k=1}^K \mathbb{E}\left[X_i^A (X_i^A)^T \middle| \mathbf{X}^B, Z_i = k\right] \mathbb{P}(Z_i = k|\mathbf{X}^B) \\ &\quad - \sum_{k=1}^K \sum_{l=1}^K \left(\mathbb{E}[X_i^A|\mathbf{X}^B, Z_i = k] \mathbb{E}[X_i^A|\mathbf{X}^B, Z_i = l]^T \mathbb{P}(Z_i = k|\mathbf{X}^B) \mathbb{P}(Z_i = l|\mathbf{X}^B) \right). \end{aligned}$$

2.2 Multivariate normal inverse Gaussian distribution

In order to achieve a model that is more flexible than a GMM, the Gaussian class distributions can be exchanged to something more flexible. In this work a multivariate generalization of the NIG distribution is used. Other mixture distributions with similar advantages have been proposed before, for instance the skewed t- and skewed normal-distributions [21, and the references within].

The probability density function of the NIG distribution is

$$f(\mathbf{x}) = \frac{\sqrt{\tau|Q|}}{(2\pi)^{(d+1)/2}} \exp\left((\mathbf{x} - \boldsymbol{\mu})^T Q \boldsymbol{\gamma} + \sqrt{2\tau}\right) 2K_\nu(\sqrt{ab}) \left(\frac{b}{a}\right)^{\frac{\nu}{2}},$$

where K_ν is a modified Bessel function of the second kind, $\nu = -\frac{d+1}{2}$, $a = \boldsymbol{\gamma}^T Q \boldsymbol{\gamma} + 2$, $b = (\mathbf{x} - \boldsymbol{\mu})^T Q (\mathbf{x} - \boldsymbol{\mu}) + \tau$. Here, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, Q and τ are parameters of the distribution where $\boldsymbol{\mu}$ is a d -dimensional location parameter, Q is a $d \times d$ -dimensional positive definite symmetric matrix defining the interdependence between the dimensions, τ is a positive scalar that parametrize the kurtosis and $\boldsymbol{\gamma}$ is d -dimensional skewness parameter.

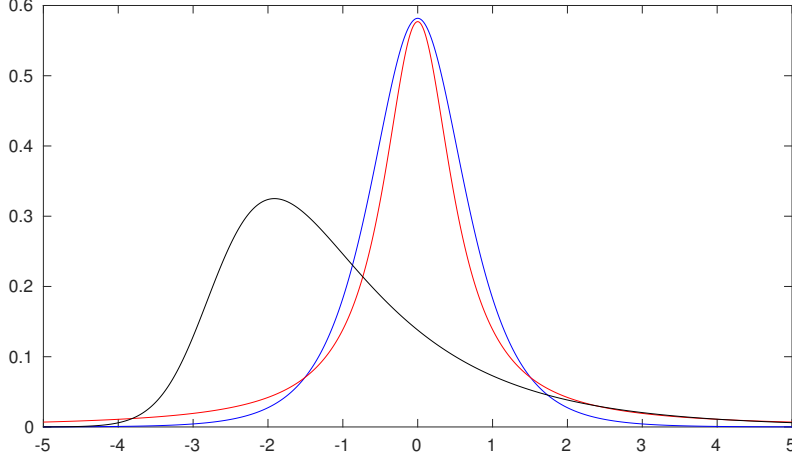


Figure 1: Examples of probability density functions for three different set of values for the parameters of a univariate NIG distribution.

A useful representation of the NIG distribution is that $\mathbf{X} \sim \text{NIG}(\boldsymbol{\mu}, Q, \boldsymbol{\gamma}, \tau)$ if

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\mu} + \boldsymbol{\gamma}V + \sqrt{V}Q^{-\frac{1}{2}}\mathbf{Z} \\ V &\sim \text{IG}(\tau, \sqrt{\frac{\tau}{2}}) \\ \mathbf{Z} &\sim \text{N}(\boldsymbol{\mu} = \mathbf{0}, \mathbf{I})\end{aligned}\tag{2}$$

Where IG denotes the inverse Gaussian distribution, see D.1, and N denotes the multivariate Gaussian distribution. Note that $\mathbf{X}|V$ is multivariate Gaussian with $\mathbb{E}[\mathbf{X}|V] = \boldsymbol{\mu} + \boldsymbol{\gamma}V$ and precision matrix $\frac{1}{V}Q$. In (2) the IG distribution is parametrized by only one parameter, τ in order to avoid overparametrization.

In comparison to the Gaussian distribution, NIG is more flexible since it allows for arbitrary skewness and kurtosis. Also, the Gaussian distribution can be characterized as a special limiting case of the NIG distribution. Figure 1 shows three examples of density functions for a univariate NIG distribution.

In order to perform probabilistic prediction of CT images the marginal and conditional distributions need to be used. The following two propositions are therefore useful.

Proposition 2.1. Suppose that $\mathbf{X} \sim \text{NIG}(\boldsymbol{\mu}, Q, \boldsymbol{\gamma}, \tau)$ and let $\mathbf{X} = [\mathbf{X}^A, \mathbf{X}^B]^T$, then $\mathbf{X}^B \sim \text{NIG}(\boldsymbol{\mu}^B, (\Sigma^{BB})^{-1}, \boldsymbol{\gamma}^B, \tau)$. Here $\boldsymbol{\mu}^B$, Σ^{BB} , and $\boldsymbol{\gamma}^B$ are the parts of $\boldsymbol{\mu}$, Σ , and $\boldsymbol{\gamma}$ respectively that correspond to \mathbf{X}^B .

Proposition 2.2. Suppose $\mathbf{X} \sim \text{NIG}(\boldsymbol{\mu}, Q, \boldsymbol{\gamma}, \tau)$, and let $\mathbf{X} = [\mathbf{X}^A, \mathbf{X}^B]^T$, then $\mathbf{X}^A|\mathbf{X}^B$ has density

$$f(\mathbf{x}^A|\mathbf{x}^B) = \frac{K_{\nu}(\sqrt{ab})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} \left(\frac{\hat{a}}{\hat{b}}\right)^{\hat{\nu}/2} \left(\frac{b}{a}\right)^{\nu/2} e^{(\boldsymbol{\gamma}^T Q(\mathbf{x} - \boldsymbol{\mu}) - \boldsymbol{\gamma}^B (\Sigma^{BB})^{-1}(\mathbf{x}^B - \boldsymbol{\mu}^B))},$$

where $\hat{\nu} = -\frac{|B|+1}{2}$, $|B|$ is the dimensionality of \mathbf{X}^B , $\hat{a} = (\boldsymbol{\gamma}^B)^T (\Sigma^{BB})^{-1} \boldsymbol{\gamma}^B + 2$ and $\hat{b} = (\mathbf{x}^B - \boldsymbol{\mu}^B)^T (\Sigma^{BB})^{-1} (\mathbf{x}^B - \boldsymbol{\mu}^B) + \tau$. Further, the conditional mean and covariance

are

$$\begin{aligned}\mathbb{E}[\mathbf{X}^A|\mathbf{X}^B] &= \boldsymbol{\mu}^A - (Q^{AA})^{-1} Q^{AB} (\mathbf{x}^B - \boldsymbol{\mu}^B) \\ &\quad + \left(\gamma^A + (Q^{AA})^{-1} Q^{AB} \gamma^B \right) \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})}, \\ \mathbb{C}(\mathbf{X}^A|\mathbf{X}^B) &= Q_{A,A}^{-1} \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} \\ &\quad + \hat{\gamma} \hat{\gamma}^T \frac{\hat{b}}{\hat{a}} \left[\left(2 \frac{\hat{\nu}+1}{\sqrt{\hat{a}\hat{b}}} - \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} \right) \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} + 1 \right].\end{aligned}$$

For a proof of these propositions, see C. Note that $\mathbf{X}^A|\mathbf{X}^B$ is not NIG distributed but has as a generalized hyperbolic distribution [28]. The generalized hyperbolic distribution generalizes NIG by letting the mixing variable, V , be distributed as a generalized inverse Gaussian distribution (GIG), see D.1 for a definition.

2.3 Spatial dependency

In a structured image, such as an MR- or CT- image, the voxel values will not be independent. While still utilizing a mixture model, we infer spatial dependency by applying a spatially dependent prior on the class membership field, $\mathbf{Z} = \{Z_i\}_i$, where Z_i indicates the class membership of voxel i . However, to simplify estimation, we still assume conditional independence between voxel values conditioned on \mathbf{Z} , i.e. $\mathbf{X}_i \perp \mathbf{X}_j | Z_i, Z_j$ for $i \neq j$.

The dependency structure in \mathbf{Z} is modeled with an MRF on the three-dimensional lattice defined through the conditional probability in (3),

$$\mathbb{P}(Z_i = k | \mathbf{Z}_{-i} = \mathbf{z}_{-i}) = \frac{1}{W_i(\mathbf{Z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp \left(-\alpha_k - \sum_{j \in \mathcal{N}_i} \beta_{kz_j} \right), \quad (3)$$

where \mathcal{N}_i is the set of all neighbors to the i :th voxel, α_k is the unconditional probability potential of class k and β_{kl} is the conditional probability potential of class k attributed to neighbors of class l . Further, $-i$ is used to denote the set of indices to all voxels except i , i.e. $-i = \{1, \dots, N\} \setminus \{i\}$ and $\mathbf{Z}_{-i} = \{Z_j\}_{j \in -i}$. Finally $W_i(\mathbf{Z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_k \exp \left(-\alpha_k - \sum_{j \in \mathcal{N}_i} \beta_{kz_j} \right)$ is a normalizing constant.

The probability density function of \mathbf{X}_i conditioned on the class identities of all other voxels is $f(x_i | \mathbf{z}_{-i}) = \sum_{k=1}^K f_k(x_i) \mathbb{P}(Z_i = k | \mathbf{Z}_{-i})$. Hence, the β_{kl} parameters describe how classes attract (negative values) or repel (positive values) each other in the topological lattice space.

Since the unconditional probability potentials, α_k , overparametrizes the conditional probability model by one degree of freedom, we let $\alpha_1 = 0$ to make the model identifiable. Furthermore, we choose a first-order neighborhood structure (the six nearest neighbors in three dimensions), see Figure 2, and for simplicity we restrict the β_{kl} values to

$$\beta_{kl} = \begin{cases} 0, & k \neq l \\ \beta, & k = l \end{cases}.$$

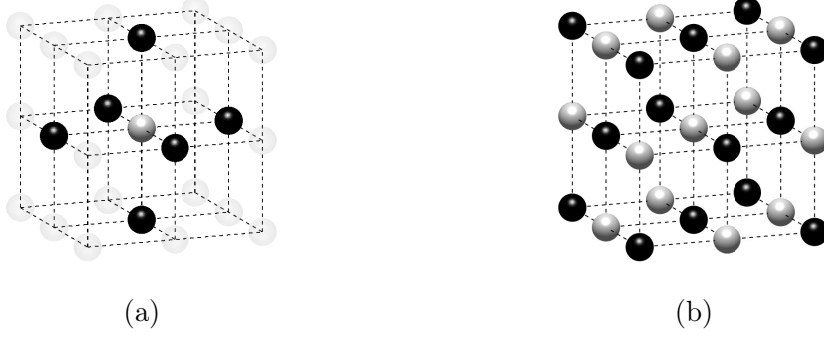


Figure 2: First-order neighborhood structure in three dimensions for an equidistant lattice. For any white ball the nearest neighbors will be black and vice versa

This means that there will only exist a conditional probability potential between neighboring voxels if they are of the same class and in this case the potential will be the same regardless of the class. This corresponds to the standard Potts model of Wu [31]. Even though we in this paper restrict β_{kl} to this simplified model, more general models with arbitrary β_{kl} values can be estimated using the same theory and methods.

From here on, we will refer to a model using an MRF prior for the class memberships as a spatial model.

3 Parameter estimation and prediction

In order to use the models in Section 2 for s-CT generation the model parameters need to be estimated from data. We choose to do this using a maximum likelihood approach. The likelihood function is

$$L(\Theta; \mathbf{x}) = \sum_{\mathbf{z} \in \Omega(\mathbf{Z})} f(\mathbf{x}|\mathbf{z}; \Theta) \mathbb{P}(\mathbf{Z} = \mathbf{z}; \Theta) = \sum_{\mathbf{z} \in \Omega(\mathbf{Z})} \left(\prod_i f(\mathbf{x}_i|z_i; \Theta) \right) \mathbb{P}(\mathbf{Z} = \mathbf{z}; \Theta),$$

where Θ is a set of parameter values for the model, \mathbf{x} is a realization of the voxel field \mathbf{X} described in Section 2.1 and $\Omega(\mathbf{Z})$ is the finite sample space of \mathbf{Z} , i.e. the set of all K^N possible combinations of class identities for the voxels.

Due to the Hammersley-Clifford theorem we know that our MRF is a neighbor Gibbs field for the first order neighborhood structure [30, Chapter 4]. Hence, there exists a closed form expression,

$$\mathbb{P}(\mathbf{Z} = \mathbf{z}) = \frac{1}{W(\boldsymbol{\alpha}, \beta)} \exp \left(- \sum_{i=1}^N \left(\alpha_{z_i} + \frac{1}{2} \sum_{j \in \mathcal{N}_i} \mathbb{I}_{z_i=z_j} \beta \right) \right). \quad (4)$$

For the spatial models, the partition function, $W(\boldsymbol{\alpha}, \beta)$, is unfortunately not feasible to compute since it requires summation over all possible states of \mathbf{Z} . Instead we replace L with the pseudolikelihood

$$\tilde{L}(\Theta; \mathbf{x}) = \sum_{\mathbf{z} \in \Omega(\mathbf{Z})} \prod_{i=1}^N f(\mathbf{x}_i|z_i; \Theta) \mathbb{P}(Z_i = z_i | \mathbf{Z}_{-i} = \mathbf{z}_{-i}; \Theta), \quad (5)$$

i.e. the joint probability of \mathbf{Z} is approximated as the product of all conditional probabilities.

For a non-spatial model the pseudolikelihood is not an approximation since the class membership of the voxels are independent of each other. For a spatial model there is however a discrepancy and approximating the joint distribution of \mathbf{Z} as in (5) can be motivated using the reasoning of [4, Section 6.1].

3.1 EM gradient algorithm

Commonly, maximum likelihood estimates of mixture models are acquired using the EM algorithm [9]. This corresponds to iteratively finding $\Theta^{(j+1)} = \arg \max Q(\Theta|\Theta^{(j)})$ where $\Theta^{(j)}$ is a vector of the estimated parameter values at the j :th iteration and

$$Q(\Theta|\Theta^{(j)}) = \mathbb{E}_{\mathbf{Z}} [\log L(\Theta|\mathbf{Z}, \mathbf{X} = \mathbf{x}) | \mathbf{X} = \mathbf{x}; \Theta^{(j)}].$$

Performing the E-step corresponds to computing the posterior probabilities for each voxels class membership, i.e. $\{\mathbb{P}(Z_i = k | \mathbf{X} = \mathbf{x}; \Theta^{(j)})\}_{i=1, k=1}^{N, K}$. For the non-spatial models, conditional class probabilities are simply $\mathbb{P}(Z_i = k | \mathbf{X} = \mathbf{x}, \Theta) = f_k(\mathbf{x})\pi_k / (\sum_l f_l(\mathbf{x})\pi_l)$. For the spatial models, only the conditional probabilities $\mathbb{P}(Z_i = k | \mathbf{z}_{-i})$ are known explicitly. However, the probability $\mathbb{P}(Z_i = k | \mathbf{x})$ can be estimated through Monte Carlo simulation since

$$\mathbb{P}(Z_i = k | \mathbf{X}; \Theta^{(j)}) = \mathbb{E}_{\mathbf{Z}} [\mathbb{I}_{Z_i=k} | \mathbf{X}; \Theta^{(j)}].$$

Here, Gibbs sampling can be used to estimate the expectation, see A for details.

Performing the M-step is straightforward for a GMM, but in general it is often difficult to derive explicit expressions for the updates. In particular, it is not computationally feasible to estimate the spatial models using a standard EM algorithm. Using an EM algorithm with an iterated conditional modes (ICM) or some Markov Chain Monte Carlo (MCMC) based estimator for the spatial parameters would be a possibility and such methods have been used in similar applications [14, 32]. However, the computational burden increases significantly if iterative methods are used in each M-step and for the purpose of whole-brain s-CT generation we need a more computationally efficient estimation method.

If it is possible to evaluate the gradient of the log likelihood with regards to the parameters, $\nabla \log \tilde{L}(\theta; \mathbf{x})$, an alternative to the EM algorithm would be to use gradient-based optimization in order to maximize the likelihood. The gradient can be expressed as

$$\begin{aligned} \nabla \log L(\Theta; \mathbf{x}) &= \nabla \log f(\mathbf{x}; \Theta) = \frac{\nabla f(\mathbf{x}; \Theta)}{f(\mathbf{x}; \Theta)} = \frac{1}{f(\mathbf{x}; \Theta)} \sum_{\mathbf{z}} \nabla f(\mathbf{x}, \mathbf{z}; \Theta) \\ &= \sum_{\mathbf{z}} \frac{f(\mathbf{x}, \mathbf{z}; \Theta)}{f(\mathbf{x}; \Theta)} \nabla \log f(\mathbf{x}, \mathbf{z}; \Theta) = \sum_{\mathbf{z}} f(\mathbf{z} | \mathbf{x}; \Theta) \nabla \log f(\mathbf{x}, \mathbf{z}; \Theta) \\ &= \mathbb{E}_{\mathbf{Z}} [\nabla \log f(\mathbf{x}, \mathbf{Z}; \Theta) | \mathbf{X} = \mathbf{x}; \Theta] \\ &= \mathbb{E}_{\mathbf{Z}} [(\nabla \log f(\mathbf{x} | \mathbf{Z}; \Theta) + \nabla \log \mathbb{P}(\mathbf{Z} = \mathbf{z}; \Theta)) | \mathbf{X} = \mathbf{x}; \Theta]. \end{aligned}$$

Analogously, for the pseudolikelihood this yields

$$\begin{aligned}\nabla \log \tilde{L}(\Theta; \mathbf{x}) &= \sum_{i=1}^N \tilde{\mathbb{E}}_{\mathbf{Z}} [\nabla \log f(\mathbf{x}_i | Z_i; \Theta) | \mathbf{X} = \mathbf{x}; \Theta] \\ &+ \sum_{i=1}^N \tilde{\mathbb{E}}_{\mathbf{Z}} [\nabla \log \mathbb{P}(Z_i = z_i | \mathbf{Z}_{-i} = \mathbf{z}_{-i}; \Theta) | \mathbf{X} = \mathbf{x}; \Theta].\end{aligned}$$

Where $\tilde{\mathbb{E}}$ denotes the expectation taken according to the probability distribution of $\mathbf{Z} | \mathbf{X}$ induced by the pseudolikelihood. The expressions inside the expectation are all available in explicit form and it is possible to estimate the expectation using Gibbs sampling in the same way as was done for $\mathbb{P}(Z_i = k | \mathbf{x})$, see A for details.

Since we can obtain an approximation of the gradient of the log likelihood, we can apply a gradient ascent algorithm to estimate the maximum likelihood parameters of the model, and iteratively update Θ as follows until convergence,

$$\Theta^{(j+1)} = \Theta^{(j)} + \delta^{(j)} \nabla \log \tilde{L}(\Theta^{(j)}; \mathbf{x}).$$

In this case, if $\delta^{(j)}$ is a sequence of step lengths with sufficiently small but positive values, $\{\Theta^{(j)}\}$ will converge to a stationary point of $\tilde{L}(\Theta; \mathbf{x})$ if one exists and if \tilde{L} is first order continuous and bounded.

Note that it is possible to evaluate the gradient if one can evaluate the conditional class probabilities, $\mathbb{P}(Z_i = k | \mathbf{X} = \mathbf{x}; \Theta)$. Finding class probabilities is equivalent to finding the expected value of the latent variable $\hat{z}_{ik} = \mathbb{I}_{Z_i=k}$ and is hence equivalent to an E-step in the regular EM algorithm. Because of this, using these gradient-based methods corresponds to, in each iteration, performing an E-step followed by taking a step in parameter space to a new set of parameter values. Thus, the method can be viewed as an EM algorithm where the M-step is approximated by one step of a gradient-based optimization method.

It is generally hard to chose values of $\delta^{(j)}$ that lead to fast and reliable convergence. Moreover, choosing the parameter path of steepest gradient is often suboptimal. One option is to replace $\delta^{(j)}$ with some scaling matrix $S^{(j)}$ that leads to a more optimal parameter path. If \tilde{L} is two times differentiable, a particular choice of $S^{(j)}$ is minus the inverse Hessian of the log likelihood. This choice of the $S^{(j)}$ matrix corresponds to Newtons method for finding zeros of $\nabla \tilde{L}$, and an update then looks like

$$\Theta^{(j+1)} = \Theta^{(j)} - H^{-1}(\log \tilde{L}(\Theta^{(j)}; \mathbf{x})) \nabla \log \tilde{L}(\Theta^{(j)}; \mathbf{x}).$$

Newtons method has superlinear convergence rate in a concave neighborhood to a stationary point [3]. This in comparison to the linear convergence rate of general choices of positive definite $S^{(j)}$ matrices (with small enough eigenvalues) suggests that Newtons method should be used when applicable.

This particular algorithm where the approximate M-step is performed by one iteration of Newtons method is known as the EM gradient algorithm [19]. This is the main outline of our estimation method. However, some modifications are needed to ensure convergence for our problem. These modifications are presented briefly in the subsections below, and the resulting estimation method is outlined in Algorithm 1.

Algorithm 1 Parameter estimation procedure.

```
1: procedure ESTIMATEPARAMS( $\mathbf{x}, \mathbf{z}, \Theta$ )
2:   while step sizes large enough do
3:     for  $k$  in  $K$  do
4:        $\mathbf{p}_k = \alpha_k + \log f(\mathbf{x}; \Theta_k)$ 
5:     end for
6:      $\{\mathbf{p}, \mathbf{z}, \mathbf{da}, \mathbf{db}\} = \text{GibbsSample}(\mathbf{x}, \mathbf{z}, \mathbf{p})$  ▷ See A
7:      $\mathbf{d}\Theta = \text{ComputeGrad}(\mathbf{x}, \mathbf{p}, \Theta)$  ▷ See section 3.1
8:      $\mathbf{s} = \text{EMGrad}(\Theta, \mathbf{d}\Theta, \mathbf{da}, \mathbf{db})$ 
9:      $\mathbf{s} = \text{LineSearch}(\Theta, \mathbf{s})$  ▷ See B
10:     $\Theta = \Theta + \mathbf{s}$ 
11:  end while
12: end procedure
```

3.1.1 Line search

Given that the Hessian matrix is negative definite for all iterations, choosing S as the negative inverse Hessian matrix scaled by some small enough step length will lead to convergence of the parameter estimation. Instead of choosing a fixed scaling of the Hessian, one could do a line search in the direction given by Newtons method to find the best scaling in each iteration. This line search procedure corresponds to performing an improved approximate M-step in the EM-gradient method and is recommended in Lange [19]. In fact, if the Hessian is negative definite and if a line search is performed, the EM gradient algorithm has a convergence rate of the same order as the regular EM algorithm.

Since we do not have a closed-form expression for the likelihood, performing line search is not straightforward for the spatial models. Instead one can perform a line search conditioned on the E-step as explained in B.

3.1.2 Sampling and conditioning of S

A problem with using a scaled negative inverse Hessian as S is that it is not guaranteed that the Hessian will be negative definite. When this assumption fails, Newtons method does not necessarily converge to a point in the parameter space with a higher likelihood than the initial value. However, by conditioning S to be positive definite, and performing a line search, it is easy to see that $Q(\Theta^{(j+1)}|\Theta^{(j)}) \geq Q(\Theta^{(j)}|\Theta^{(j)})$, and the inequality will be strict as long as $\Theta^{(j)}$ is not a stationary point. To achieve a positive definite S we first check if the computed Hessian is negative definite. If not, S is chosen as the negative diagonal of the Hessian. If S is still not positive definite, the diagonal entries are translated until all of them are positive.

A method that iteratively performs an E-step followed by an approximate M-step that guarantees $Q(\Theta^{(j+1)}|\Theta^{(j)}) > Q(\Theta^{(j)}|\Theta^{(j)})$ is known as a Generalized EM algorithms (GEM) [9]. Thus, by conditioning S to be positive definite and performing a line search conditioned on the E-step, the proposed estimation algorithm belongs to the class of GEMs. Hence, it will always increase the likelihood and when the assumptions for Newtons method are fulfilled it will also converge with super linear rate.

The E-step for the spatial models is only approximate due to the MC sampling. So far

we have not assessed if the Monte Carlo errors affects the convergence results. However, in analogy to the MCEM algorithm of Wei and Tanner [29], the MCMC approximation in our E-step give us, not a GEM method, but a Monte Carlo GEM method. Convergence of such an algorithm follow analogously from the convergence of the MCEM algorithm [8].

3.1.3 Approximate Hessian

The final modification that is needed to make the estimation method work is to approximate the Hessian. The reason for this is that it is not possible to estimate the true Hessian using Monte Carlo simulation in the same manner as was possible with the gradient. The Hessian can be written as

$$\begin{aligned} H(\log \tilde{L}(\Theta; \mathbf{x})) &= \sum_{i=1}^N \tilde{\mathbb{E}}_{\mathbf{Z}} [H(\log f(\mathbf{x}_i|Z_i; \Theta)) + H(\log f(Z_i|\mathbf{Z}_{-i}; \Theta)) | \mathbf{X} = \mathbf{x}; \Theta] \\ &+ \sum_{i=1}^N \tilde{\mathbb{E}}_{\mathbf{Z}} [\nabla \log f(\mathbf{x}_i|Z_i; \Theta) \nabla \log f(\mathbf{Z}|\mathbf{X} = \mathbf{x}; \Theta)^T | \mathbf{X} = \mathbf{x}; \Theta] \\ &+ \sum_{i=1}^N \tilde{\mathbb{E}}_{\mathbf{Z}} [\nabla \log f(Z_i|\mathbf{Z}_{-i}; \Theta) \nabla \log f(\mathbf{Z}|\mathbf{X} = \mathbf{x}; \Theta)^T | \mathbf{X} = \mathbf{x}; \Theta]. \end{aligned}$$

Here, the last two terms include $\nabla \log f(\mathbf{Z}|\mathbf{X} = \mathbf{x}; \Theta)^T$ which we do not have a closed form expression for. Instead, we will approximate the Hessian using only the first term, in hope that this term dominates the Hessian. This approximation highlight yet another reason why it is necessary for us to perform a proper line search and conditioning of S to assure convergence.

3.2 CT prediction

Given parameter estimates based on training data, the conditional expectation from Equation (1) can be used to generate the s-CT images for a new patient where only MR images are recorded. Thus, the CT value for each voxel is predicted using the formula

$$\mathbb{E}[X_i^A | \mathbf{X}^B = \mathbf{x}^B] = \sum_{k=1}^K \mathbb{E}_k[X_i^A | \mathbf{X}_i^B = \mathbf{x}_i^B] \mathbb{P}(Z_i = k | \mathbf{X}^B = \mathbf{x}^B),$$

where X_i^A is the CT value of the i :th voxel and \mathbf{X}_i^B are the MR values for the same. Here, $\mathbb{E}_k[X_i^A | \mathbf{X}_i^B = \mathbf{x}_i^B]$ has an analytical closed form expression and $\mathbb{P}(Z_i = k | \mathbf{X}^B = \mathbf{x}^B)$ can be approximated using MCMC simulation analogously to how $\mathbb{P}(Z_i = k | \mathbf{X} = \mathbf{x})$ was approximated, see A.

Using the conditional mean as the predictor of the s-CT image corresponds to minimizing the root mean square error (RMSE) of the prediction, based on the model assumption being correct. If one instead would be interested in minimizing the mean absolute error (MAE) the conditional median would be a more appropriate predictor [11]. There is no analytical expression for the conditional median but given $\mathbb{P}(Z_i = k | \mathbf{X}^B = \mathbf{x}^B)$ it can easily be approximated by Monte-Carlo simulation since the conditional distribution is known.

3.3 Computational cost

In A it is shown that the computational complexity of the MCMC sampling used to approximate the expectations in the EM gradient method is of order $\mathcal{O}(JNK)$, where J is the number of Monte Carlo iterations, K is the number of classes and N is the number of voxels in the image. Here, J can be chosen low since MCMC chains of consecutive EM-gradient iterations can feed of the former to reduce burnin. For the results presented later, $J = 10$ was found to be sufficient. Besides the MCMC sampling, each iteration of the algorithm requires summing up computations voxelwise as well as classwise. Hence, the computational complexity is $\mathcal{O}(nJNK)$, where n is the number of EM gradient iterations.

Analogously, once the model parameters are available, the CT prediction has a computational complexity of $\mathcal{O}(JNK)$ due to the need to generate a MCMC chain for approximating $\mathbb{P}(Z_i = k | \mathbf{X}^B = \mathbf{x}^B)$. For the prediction step, J needs to be larger in order to get rid of the burn in phase since there are no consecutive iterations to feed from. However, this is not a big issue since the prediction step only is performed once, not iteratively, and the number of voxels of one CT image are smaller than that of all the voxels from all images used in the training set. For the prediction, $J = 1000$ was found to be sufficient.

The important thing to note here is that, both for parameter estimation and CT prediction, the scaling in N is linear which makes it feasible to fit the spatial models to large data sets such as multiple whole-brain images.

4 An application to real data

Estimation of CT like images from MR data can be done in a great variety of ways, using for example atlases [1, 10], or segmentation based [15] or combinations [25]. At present there are no way to directly compare these methods in terms of accuracy as they are based on different types of MR sequences acquired at different MR scanners with different field strength and with different coil solutions. The different studies report results for different areas of the body and have different inclusion and exclusion criteria's. To make a meaningful comparison of the proposed method we compare it to the previously published method of Johansson et al. [16] (GMM), using the same input data.

This section presents the result from such a comparative cross-validation study. The models evaluated are:

1. Gaussian mixture model with spatially independence (GMM).
2. Gaussian mixture model with spatially dependence (GMMS).
3. NIG mixture model with spatially independence (NIG).
4. NIG mixture model with spatially dependence (NIGS).

Here the spatial dependence refer to the spatial prior of Section 2.3.

The data, described further in Section 4.1, is three-dimensional images from scans of 14 patients. The results were analyzed using leave-one-out cross-validation with one fold for each patient. For each fold, parameters were estimated using the parameter estimation method described in Section 3 and data from all but but that fold. S-CT images was then

generated for the fold using the estimated parameters and the MR images for the fold. Two s-CT images were generated for each model. One using the conditional expectation and one using the conditional median. The differences between the true CT images and the generated s-CT images were compared using MAE and RMSE. MAE is here chosen as the main metric for performance assessment since the amount of radiation that is attenuated on the way through a body is proportional to the accumulation of CT values over the radiation path. Hence, the conditional median should be the correct predictor to use for s-CT generation in order to assess the models [11].

The model's ability to explain the distribution of the CT values conditioned on the MR images was also evaluated using the negatively oriented CRPS score [12], see E.

Choosing the parameter K , the number of classes, is not part of the estimation method. Hence, we are evaluating all the models for mixture classes ranging from 2 up to 10 in order to assess the sensitivity of this parameter.

Since neither the EM- nor the EM gradient- algorithm necessarily finds a global optima it is common to run the parameter estimation procedure several times with different and randomized initial values. In our implementation, the GMM was initialized using 15 randomized starting values as well as two starting values acquired from estimating each class parameters from the data associated to it using a kmeans- and a hierarchical-clustering algorithm [13]. For the other three models (GMMS, NIG and NIGS), only two initial values were used. One with the initial values derived from the estimated parameters of the GMM and one using just a set of constant values such as for instance a diagonal of ones as the precision matrices. Model selection among the initial values were chosen on the basis of lowest MAE of CT predictions on the training data. Performing model selection like this is better suited for our particular problem of s-CT generation compared to choosing the model with the highest likelihood.

The main results can be seen in Section 4.2 and some auxiliary results in Section 4.3.

4.1 Details about data

The data used in this study consists of images from 14 different patients which were included in the study after oral and written consent. The study has been approved by the regional ethical review board of Umeå University. For each patient, one CT image and four MR images were acquired. For the MR images, two dual echo UTE sequences were used, one with a flip angle of 10° and one with a flip angle of 30° . The UTE sequences sampled a first echo at 0.07 ms and a second echo at 3.76 ms. For both sequences the repetition time was 6 ms. Two different flip angles and two different echo times give four possible combinations and hence four different MR images.

The difference between images from the first and second echo indicate the presence of short $T2^*$ tissues. The differences between images acquired using the two flip angles indicate presence of $T1$ tissues. A short $T2^*$ is not only found in tissues with a short $T2$, but also in regions with rapid coherent dephasing, such as air-soft tissue and bone-soft tissue interfaces. Knowing the $T1$ information can help to separate these interfaces from $T2$. This is of interest since the $T2$ value is a good discriminator between bone, soft tissue and air [16].

All MR images were acquired with a 1.5 T Siemens Espree scanner. The UTE images were reconstructed to $192 \times 192 \times 192$ voxel bitmapped images with an isotropic resolution

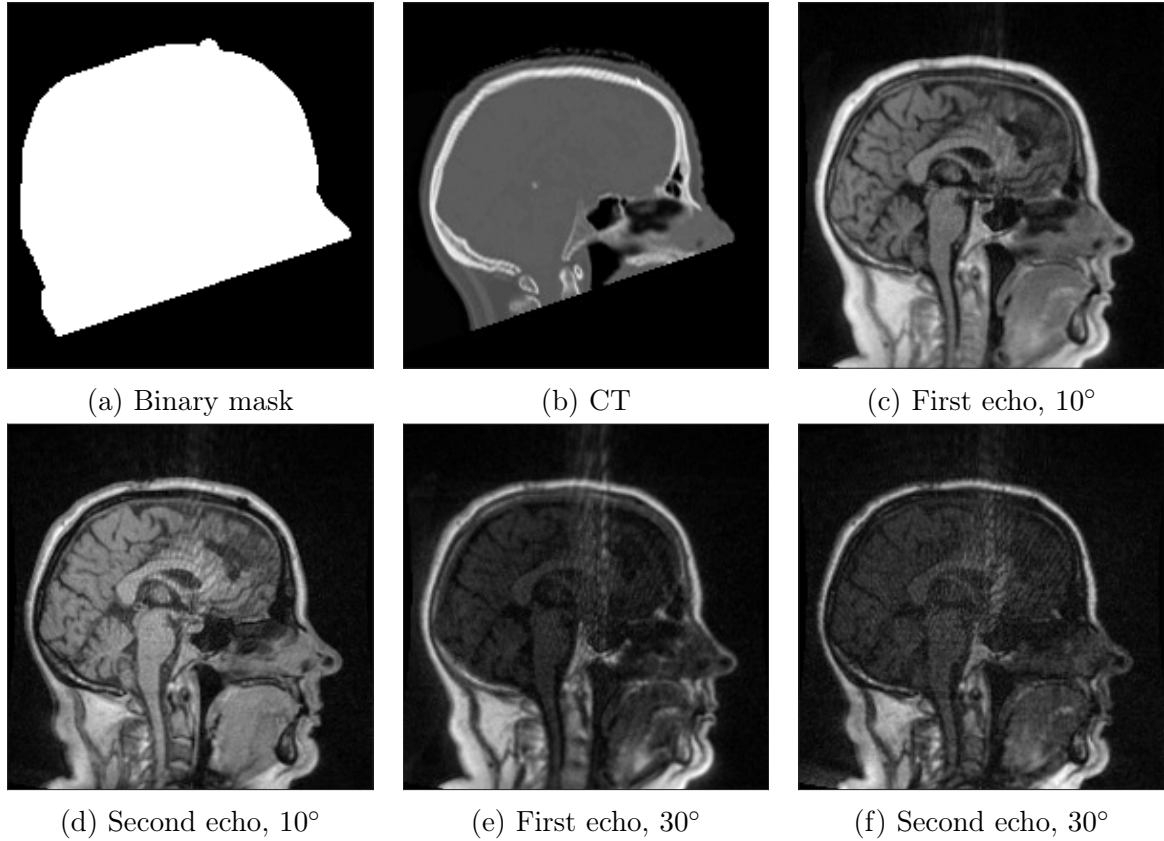


Figure 3: Binary data mask (panel a), CT image (panel b), The four MRI UTE sequences (panels c-f).

and a voxel size of 1.33 mm. The UTE sequences sampled the k-space radially with 30 000 radial spokes. CT images were acquired with a tube voltage of between 120 kV and 130 kV on either a GE Lightspeed Plus, Siemens Emotion 6 or GE Discovery 690. The in-plane pixel size varied between 0.48 mm to 1.36 mm and the slice thickness between 2.5 mm and 3.75 mm. Images of the same patient were co-registered and resampled to achieve voxel-wise correspondence between all five modes. A binary mask excluding most of the air surrounding the head was computed from the images and used to remove unnecessary data. Furthermore, to reduce the execution time of the parameter estimation phase, only 11 slices in the middle of the head of each patient was used during the parameter estimation phase, but all slices were used during the prediction phase (s-CT generation). Additional details concerning the data can be found in Johansson et al. [16]. Data from one slice of a patient is shown in Figure 3.

Figure 4 shows a smoothed histogram of the CT values for all patients. Note that most voxels have CT values close to zero HU. This corresponds to soft tissue which makes up the main volume of the head. The peak at around -1000 HU in the histogram corresponds to air. The presence of air is partly due to the cavities in the nasal region, sinuses and throat but also partly due to that the binary mask around the head is not completely tight and allow air in between the actual outline of the head and what the mask cuts away. The higher CT values (typically around 600 to 1500) corresponds to bone.

The very high CT values (> 1500) correspond to streak artifacts or interpolation

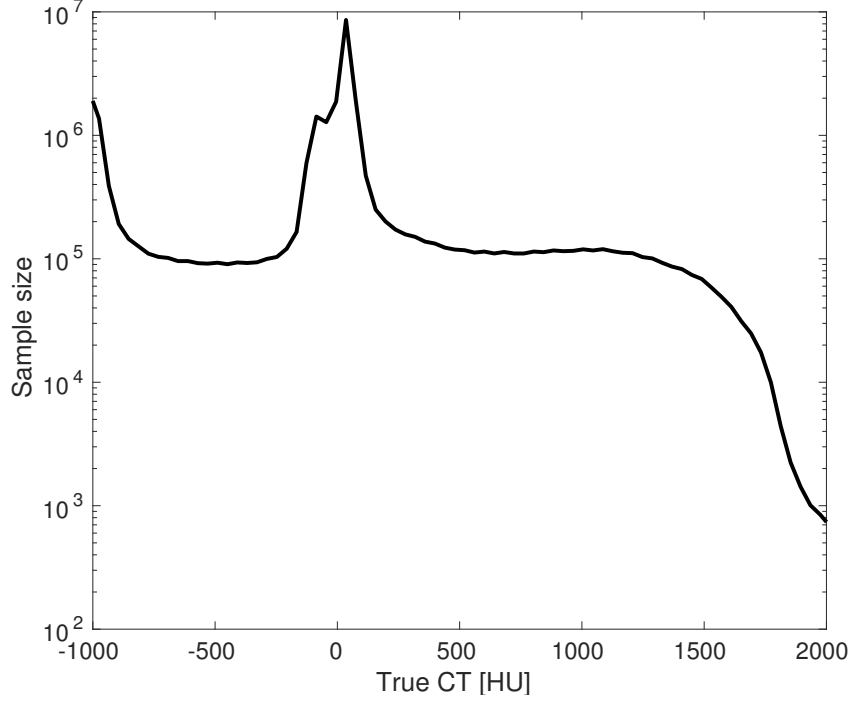


Figure 4: Smoothed histogram of CT values in the total data set of all 14 patients. Note the logarithmic scale on the y-axis.

errors in the resampling procedure.

4.2 Main results

The cross-validation study was carried out for each of the four models described above, and each model was tested with the number of mixture classes ranging from two up to ten. MAE and RMSE values from the study can be seen in Figure 5 and in Table 1. The lowest MAE value (146.4 HU) was attained using the NIGS model with seven classes and conditional mean. Except for two- or three- class models the NIGS model followed by the GMMS had the lowest prediction error both in MAE and RMSE.

The conditional median improves the MAE for the non-spatial models but do not affect the spatial ones as clearly. For the GMMS there is almost no difference between using mean or median as predictor. For the NIGS it improves MAE when using two to five classes but performs similarly when $K > 5$. For RMSE it is a consistent drawback of using median instead of mean as the predictor, as expected.

Figure 6 shows prediction errors as functions of predicted CT values. Comparing the panels one can note that the prediction errors for all models are smaller for voxels where the predicted CT value is in an interval with more frequent CT values. Here this corresponds to soft tissue (around 0 HU) and air (around -1000 HU).

Figure 6c shows the bias of the predictions. Here, the spatial models seem to have a generally bigger bias. However, at the same time the spatial models have a generally lower MAE and RMSE, especially when predicting CT values above 0, i.e. bone. This suggests that the variance of the estimates are smaller for the spatial models.

The negatively oriented CRPS (from now on referred to as the CRPS*) is a measure

Table 1: Prediction errors of the models at the number of classes were MAE reached its minimum for each model. The models are compared both using mean and median as predictor. "Ratios" compare the corresponding error with the reference model (GMM with mean predictor).

Model	Predictor	Classes	MAE	MAE Ratio	RMSE	RMSE Ratio
GMM	Mean	9	178.3	100%	353.4	100%
GMM	Median	9	171.5	96.2%	373.4	105.7%
GMMS	Mean	10	154.9	86.9%	325.1	92.0%
GMMS	Median	10	155.8	87.4%	330.0	93.4%
NIG	Mean	7	179.4	100.5%	351.7	99.5%
NIG	Median	7	168.1	94.3%	370.0	104.7%
NIGS	Mean	7	146.4	82.1%	308.5	87.3%
NIGS	Median	7	147.3	82.6%	316.7	89.6%

of how well a probability distribution explains the observed data, see E for further details. A small CRPS* indicates a good distributional fit. Note that, CRPS* is only associated with the conditional model and not with the chosen prediction function derived from it. Figure 7 show the mean CRPS* over all predicted voxels for each model and number of classes. As can be seen, the CRPS follow the same behavior as the errors of Figures 5 except that the NIG model performs worse than the GMM.

Figure 8 shows the true CT image together with the corresponding s-CT images for a selected patient. Since the data is three-dimensional we present it as an intersection viewed in profile of the head. The absolute errors and conditional standard deviation are also shown for comparison. Note that the non-spatial models give more noisy predictions compared to the spatial ones. All models seem to have the most trouble predicting values in the regions where air, soft tissue and bone interacts such as in the nasal and throat cavities. These regions have short T2* values even without presence of low T2 values due to susceptibility effect [16], and apparently the T1 information acquired by the two flip angles is not enough to classify these regions correctly. For the spatial model these prediction problems are amplified by the fact that there are many small regions of soft tissue, air and bone close to each other. Since the spatial model allow for classes to cluster by spatial attraction, this will diminish the probability of classifying small regions close to each other to different classes as would be needed to make proper predictions in these regions.

Apparently this clustering effect of the spatial models are advantageous overall, since otherwise the estimated models would have had β values close to zero.

4.3 Median filtering

The purpose of the spatially dependent model is to increase the probability of a certain class for a voxel when this class is predominately common in the neighborhood of said voxel. This in turn will make regions of s-CT images more homogeneous while still allowing for sharp edges in between two tissue types.

Applying a two-dimensional median filter (letting each voxel assume the median value in a defined spatial neighborhood to it) [27] in a post-processing step to a non-spatial

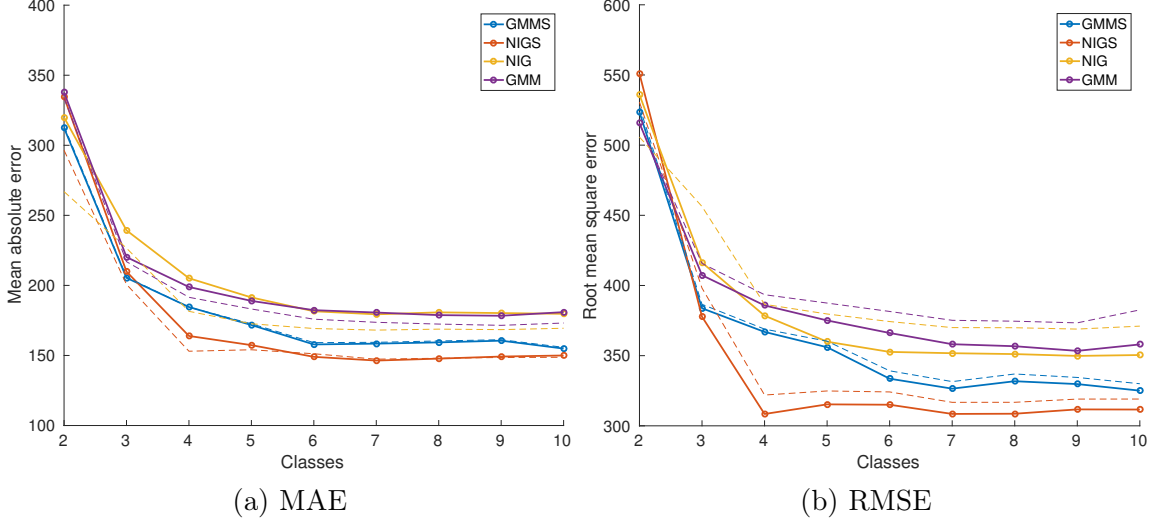


Figure 5: Errors of s-CT images compared to the number of classes in the mixture model. Errors shown for both conditional mean (thick solid lines) and conditional median predictions (thin dashed lines).

model would be one simple but less controlled way of giving the images these properties. A natural question is if our non-spatial model with such a median filter is comparable in prediction performance to the more complicated spatial models. In order to assess this we applied a median filter with a kernel of radius 2 (each voxel is the median of the voxel values among the nearest neighbors including the voxel itself) to the predicted images before calculating the prediction errors. This post-processing step was performed for all of the four models and Table 2 show the corresponding prediction errors.

One can conclude that median filtering improves the s-CT images both in MAE and RMSE sense. For the non-spatial models the errors are reduced with about 5% and for the spatial models with about 2% in MAE (similar improvements for RMSE). Even after the median filtering step the spatial models show a considerable advantage over the non-spatial ones.

5 Discussion

We have presented a class of spatially dependent mixture models that can be used to generate a three dimensional substitute CT image using information from MR images. We also introduced a computationally efficient algorithm that can perform maximum likelihood parameter estimation of the model without the need for ever evaluating the actual likelihood. This estimation method is applicable to a much larger class of problems where the likelihood is intractable to compute but where the gradient of it can be approximated.

The proposed model (NIGS) and variations of it were compared with a reference model (GMM) that has already shown promising research results. The NIGS model, which uses NIG mixture distributions and spatially dependent prior probabilities of the latent class memberships, attained the smallest prediction error measured both in MAE and RMSE. Compared to using a non-spatial GMM the prediction error decreased with 17.9% in

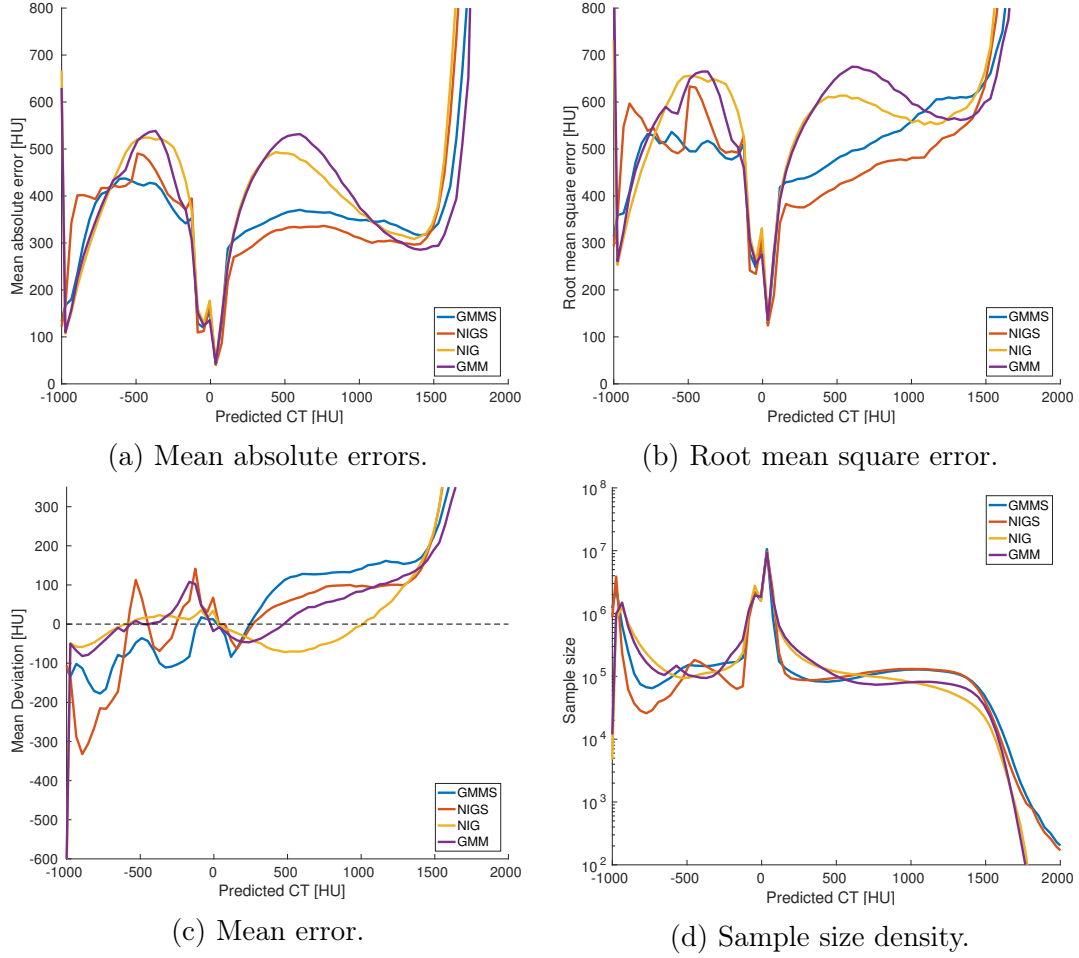


Figure 6: The MAE, RMSE, mean error and sample size density as functions of the predicted CT values [HU] for the four different models at the number of classes that minimized MAE.

MAE and 12.7% in RMSE. Using the spatially dependent model with Gaussian mixture distributions (GMMS) showed advantages to the non-spatial models but had larger prediction errors in both MAE and RMSE compared to NIGS. The spatially independent NIG mixture model (NIG) showed similar prediction performance to the original GMM model.

Compared to the work of Johansson et al. [16] we also evaluated the conditional median as a predictor for s-CT generation. If one is mainly interested in minimizing the MAE this is theoretically a better predictor and for a small number of mixture classes, conditional median yielded a smaller MAE than conditional mean for all models. For the NIGS model, the gain of decreasing the MAE by using the conditional median is declining with the number of classes and after five classes there is no apparent difference whichever predictor is used. For the GMMS, the MAE was comparable in all cases. The non spatial models showed a consistent advantage of using the conditional median for minimizing MAE. The RMSE was larger for all models when using the conditional median. Therefore we recommend using the conditional mean for predicting CT values when working with the spatial models.

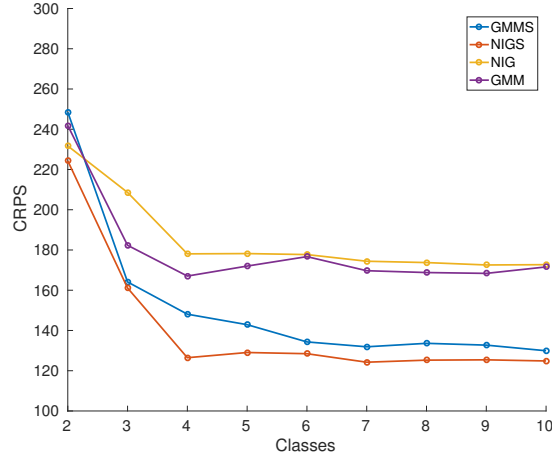


Figure 7: CRPS* for all models and number of classes.

Table 2: Prediction errors of the models at the number of classes were MAE reached its minimum for each model when using a spatial median filter on the predicted CT values in a post-processing step. The models are compared both using mean and median as predictor. "Ratios" compare the corresponding error with the reference model (GMM with mean predictor).

Model	Classes	MAE Median	MAE Ratio	RMSE Median	RMSE Ratio
GMM	9	168.2	94.3%	328.7	93.0%
GMM Median	9	160.2	89.9%	342.0	96.8%
GMMS	10	151.6	85.0%	316.8	89.6%
GMMS Median	10	152.4	85.5%	321.2	90.9%
NIG	7	170.4	95.6%	331.1	93.7%
NIG Median	7	157.0	88.1%	339.5	96.1%
NIGS	7	143.8	80.7%	302.4	85.6%
NIGS Median	7	144.5	81.0%	309.9	87.7%

The mean CRPS* values indicate that not only the point estimates but also the conditional distributions in general are more accurate for the NIGS and GMMS models. The NIG model is however consistently worse of than GMM in terms of CRPS*. This is surprising since the GMM is a limiting case of NIG, especially since it in conjunction with the spatial model clearly outperforms GMMS.

A median filtering post processing step showed a slight advantage for all models. Both MAE and RMSE errors decreased and at most it yielded a 5.7% improvement in MAE. The comparatively small gain shows that the predictive performance of the spatial model can not simply be synthesized by a median filtered non-spatial prediction. At the same time we would recommend using such a post-processing step since there are some improvements, especially to the non-spatial models. Using the NIGS model with the median filter give a 19.3% improvement compared to the reference model.

All four models exhibited a decreasing marginal gain of adding further classes and using more than seven does not affect the predictive performance significantly. Since, the computational cost increases linearly with the number of classes, the parameter estimation converges slower, and there will at some point be a risk of overfitting, we suggest choosing

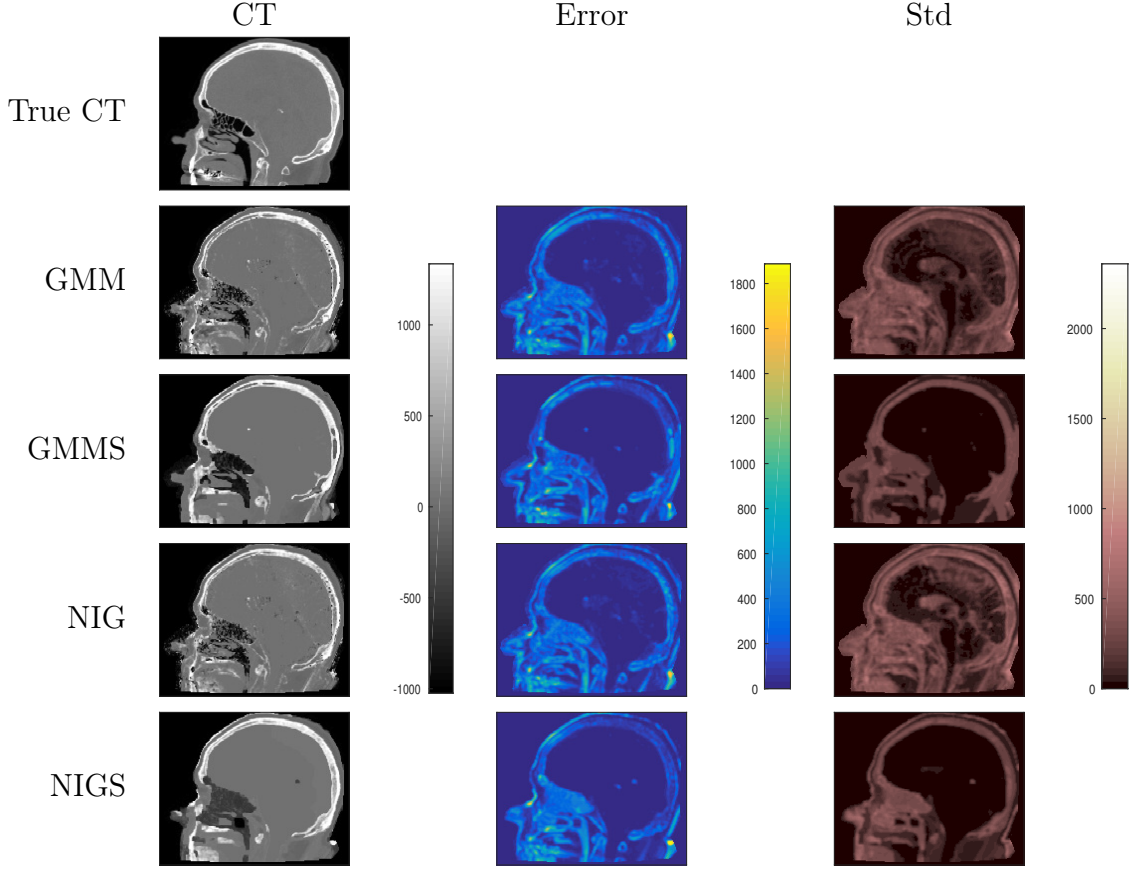


Figure 8: Generated s-CT images for one of the patients in the study. The first row shows the true CT images. The other four rows show the generated s-CT images using the four different models. The different columns show the s-CT images (left), the voxelwise absolute errors between the corresponding s-CT and true CT image (mid), and the voxelwise conditional standard deviations (right).

$K = 7$ for all of the models.

The regions with the worst predictive performance seem to be the nasal/throat cavities. This is in line with prior work [16] and is mainly due to lack of information in the MR images for the regions. The spatial model did not show any advantages in these regions and one can even argue that the model inherently counteracts enhanced prediction due to the spatially quick alternations of tissue types in these regions. If these regions are of particular importance in an application the spatial model could be enhanced by allowing for spatially varying β parameters. One could for instance let the β parameter be a low order polynomial of the spatial coordinates in some reference system, for instance the reference system proposed in Johansson et al. [17]. Estimating β would then correspond to estimating the coefficients of the polynomial. This can be done using the proposed gradient method and is something to include in future work.

Acknowledgements

This work has been supported by grants from the Knut and Alice Wallenberg foundation and the Swedish Research Council Grants 2008-5382 and 340-2013-5342.

References

- [1] H. Arabi, N. Koutsouvelis, M. Rouzaud, R. Miralbell, and H. Zaidi. Atlas-guided generation of pseudo-ct images for mri-only and hybrid pet-mri-guided radiotherapy treatment planning. *Physics in Medicine and Biology*, 61(17):6531–6552, 2016.
- [2] O. Barndorff-Nielsen. Normal inverse gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, 24(1):1–13, 1997.
- [3] Dimitri P. Bertsekas. *Nonlinear Programming*, chapter Unconstrained Optimization. Athena Scientific, 1995.
- [4] J. Besag. Interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.
- [5] I. Bezrukov, F. Mantlik, H. Schmidt, B. Schölkopf, and B.J. Pichler. MR-based PET attenuation correction for PET/MR imaging. *Seminars in nuclear medicine*, 43(1): 45–59, 2013.
- [6] D. Bolin and J. Wallin. Multivariate normal inverse Gaussian Matérn fields. *ArXiv e-prints*, 1606.08298, June 2016.
- [7] G. Celeux, F. Forbes, and N. Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36(1): 131–144, 2003.
- [8] K.S. Chan and J. Ledolter. Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252, 1995.
- [9] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [10] J.A. Dowling, J. Lambert, J. Parker, O. Salvado, J. Fripp, A. Capp, C. Wratten, J.W. Denham, and P.B. Greer. An atlas-based electron density mapping method for magnetic resonance imaging (mri)-alone treatment planning and adaptive mri-based prostate radiation therapy. *International journal of radiation oncology, biology, physics*, 83(1):e5–e11, 2012.
- [11] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.
- [12] T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

- [13] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of statistical learning*, chapter Unsupervised Learning. Springer, second edition, 2009.
- [14] K. Held, E.R. Kops, B.J. Krause, W.M. Wells, R. Kikinis, and H. Müller-Gärtner. Markov random field segmentation of brain MR images. *IEEE Transactions on medical imaging*, 16:878–886, 1997.
- [15] S. Hsu, Y. Cao, K. Huang, M. Feng, and J.M. Balter. Investigation of a method for generating synthetic ct models from mri scans of the head and neck for radiation therapy. *Medical Physics*, 58(23):8419–8435, 2013.
- [16] A. Johansson, M. Karlsson, and T. Nyholm. CT substitute derived from MRI sequences with ultrashort echo time. *Medical Physics*, 38:2708–2714, 2011.
- [17] A. Johansson, A. Garpebring, M. Karlsson, T. Asklund, and T. Nyholm. Improved quality of computed tomography substitute derived from magnetic resonance (MR) data by incorporation of spatial information. *Acta Oncologica*, 52:1369–1373, 2013.
- [18] J. Jonsson, M.M. Akhtari, M.G. Karlsson, A. Johansson, T. Asklund, and T. Nyholm. Accuracy of inverse treatment planning on substitute CT images derived from MR data for brain lesions. *Radiation Oncology*, 10:1–7, 2015.
- [19] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):425–437, 1995.
- [20] A. Larsson, A. Johansson, J. Axelsson, T. Nyholm, T. Asklund, K. Riklund, and M. Karlsson. Evaluation of an attenuation correction method for PET/MR imaging of the head based on substitute CT images. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 26:127–136, 2013.
- [21] S. Lee and G.J. McLachlan. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202, 2014.
- [22] C.P. Robert and G. Casella. *Monte Carlo statistical methods*, chapter Controlling Monte Carlo Variance. Springer, 2004.
- [23] C.P. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. Springer, 2010.
- [24] X. Rongjing and J. Neville. Pseudolikelihood em for within-network relational learning. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1103–1108, 2008.
- [25] C. Siversson, F. Nordström, T. Nilsson, T. Nyholm, J. Jonsson, A. Gunnlaugsson, and L.E. Olsson. Technical note: Mri only prostate radiotherapy planning using the statistical decomposition algorithm. *Medical Physics*, 42(10):6090–6097, 2015.
- [26] J. Sjölund, D. Forsberg, M. Andersson, and H. Knutsson. Generating patient specific pseudo-CT of the head from MR using atlas-based regression. *Physics in medicine and biology*, 60(2):825–839, 2015.

- [27] M Sonka, V Hlavac, and R Boyle. *Image Processing, Analysis, and Machine Vision*, chapter Image pre-processing. Thomson, 2008.
- [28] E.A. Von Hammerstein. *Generalized hyperbolic distributions: Theory and applications to CDO pricing*. PhD thesis, Albert-Ludwigs-Universität Freiburg, 2010.
- [29] G.C.G. Wei and M.A. Tanner. A monte carlo implementation of the EM algorithm and the poor man’s data augmentation algorithm. *Journal of the American Statistics Association*, 85(411):699–704, 1990.
- [30] Gerhard Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Springer, 2003.
- [31] F. Y. Wu. The Potts model. *Reviews of Modern Physics*, 54(1):235–268, 1982.
- [32] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

A MCMC sampling from the latent field

In Section 3.1 the posterior probability of voxel i to be a member of class k was needed to perform an E-step. These probabilities are equivalent to taking expectations of indicator functions over the latent field \mathbf{Z}_{-i} conditioned on the observed data. As was stated, we can not calculate these expectations explicitly so instead we estimate them by Monte Carlo integration. In the more general case we have an expression on the form $\mathbb{E}_{\mathbf{Z}}[g(Z_i, \mathbf{x}_i, \mathbf{Z}_{-i})|\mathbf{x}]$ for some function $g(Z_i, \mathbf{x}_i, \mathbf{Z}_{-i})$ that is explicitly available. A Monte Carlo approximation for such an expression yields

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}[g(Z_i, \mathbf{x}_i, \mathbf{Z}_{-i})|\mathbf{X} = \mathbf{x}] &= \sum_{\mathbf{z}} g(z_i, \mathbf{x}_i, \mathbf{z}_{-i}) \mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x}) \\ &\approx \frac{1}{J} \sum_{j=1}^J g\left(\hat{z}_i^{(j)}, \mathbf{x}_i, \hat{\mathbf{z}}_{-i}^{(j)}\right),\end{aligned}$$

where J is the number of realizations in the MC simulation and $\hat{\mathbf{z}}^{(j)}$ is the j :th realization of $\mathbf{Z}|\mathbf{X}$. Using Rao-Blackwellization [22] a more efficient estimation of the conditional expectation can be computed as

$$\mathbb{E}_{\mathbf{Z}}[g(Z_i, \mathbf{x}_i, \mathbf{Z}_{-i})|\mathbf{X} = \mathbf{x}] \approx \frac{1}{J} \sum_{j=1}^J \sum_{k=1}^K g\left(k, \mathbf{x}_i, \hat{\mathbf{z}}_{-i}^{(j)}\right) \mathbb{P}\left(Z_i = k \middle| \mathbf{X} = \mathbf{x}, \mathbf{Z}_{-i} = \hat{\mathbf{z}}_{-i}^{(j)}\right) \quad (6)$$

To be able to sample realizations of $\mathbf{Z}|\mathbf{X}$ one can note that by Bayes theorem $\mathbb{P}(Z_i|\mathbf{Z}_{-i}, \mathbf{X}) \propto f(\mathbf{X}_i|Z_i)\mathbb{P}(Z_i|\mathbf{Z}_{-i})$. Further, it is possible to sample from the full conditionals of $\mathbf{Z}|\mathbf{X}$ and Gibbs sampling [30, Chapter 5] can therefore be used. A blocking algorithm can be utilized to enhance the performance of the Gibbs sampling by partitioning the voxels in two mutually exclusive sets $\mathbf{Z} = [\mathbf{Z}_b, \mathbf{Z}_w]$ as in Figure 2b (black

and white balls). Due to the first-order neighborhood structure of the MRF model, the full conditional probabilities are only dependent on the nearest neighbors, i.e. the voxels of \mathbf{Z}_b are independent of each other conditioned on \mathbf{Z}_w and vice versa. A blocked Gibbs sampler can therefore be used, by iteratively sample from $\mathbf{Z}_b|\mathbf{Z}_w, \mathbf{x}$ and then $\mathbf{Z}_w|\mathbf{Z}_b, \mathbf{x}$.

The pseudolikelihood was derived by making an approximation on the joint distribution of \mathbf{Z} , see Equation (5). This approximation induces a posterior distribution $\tilde{\mathbb{P}}(\mathbf{Z}|\mathbf{X})$. To maximize the pseudolikelihood, the E-step in both the EM- and EM gradient-algorithm should be performed with regards to this induced probability distribution. The above sampling scheme samples from the true $\mathbf{Z}|\mathbf{X}$. In order to sample from the induced distribution one can construct importance weights from the following relation:

$$\tilde{\mathbb{P}}(\mathbf{Z}|\mathbf{X}) = C(\boldsymbol{\alpha}, \beta) \frac{\mathbb{P}(\mathbf{X}|\mathbf{Z})}{\mathbb{P}(\mathbf{X})} \mathbb{P}(\mathbf{Z}_b|\mathbf{Z}_w) \mathbb{P}(\mathbf{Z}_w|\mathbf{Z}_b) = C(\boldsymbol{\alpha}, \beta) \frac{\mathbb{P}(\mathbf{Z}_w|\mathbf{Z}_b)}{\mathbb{P}(\mathbf{Z}_w)} \mathbb{P}(\mathbf{Z}|\mathbf{X}),$$

where $C(\boldsymbol{\alpha}, \beta)$ is a normalizing constant for the induced probability mass function. This gives that

$$\tilde{\mathbb{E}}_{\mathbf{Z}}[g(Z_i, \mathbf{x}_i, \mathbf{Z}_{-i})|\mathbf{X} = \mathbf{x}] = \mathbb{E}_{\mathbf{Z}} \left[g(Z_i, \mathbf{x}_i, \mathbf{Z}_{-i}) C(\boldsymbol{\alpha}, \beta) \frac{\mathbb{P}(\mathbf{Z}_w|\mathbf{Z}_b)}{\mathbb{P}(\mathbf{Z}_w)} \middle| \mathbf{X} = \mathbf{x} \right].$$

By the use of the Hammersley-Clifford theorem [30, Chapter 4] the denominator of the correction factor $c(\mathbf{Z}) = C(\boldsymbol{\alpha}, \beta) \frac{\mathbb{P}(\mathbf{Z}_w|\mathbf{Z}_b)}{\mathbb{P}(\mathbf{Z}_w)}$ is known up to a normalizing constant, see equation (4). $c(\mathbf{Z})$ can therefore be expressed as

$$c(\mathbf{Z}) = C(\boldsymbol{\alpha}, \beta) W(\boldsymbol{\alpha}, \beta) \frac{\prod_{l \in w} \exp(-\sum_{m \in \mathcal{N}_l} \mathbb{I}_{z_l = z_m} \beta)}{\prod_{l \in b} \left(\sum_{k=1}^K \exp(-\alpha_k - \sum_{m \in \mathcal{N}_l} \mathbb{I}_{z_m = k} \beta) \right)}.$$

Where $W(\boldsymbol{\alpha}, \beta)$ is the unknown partition function of $\mathbb{P}(\mathbf{Z})$, w is the set of all voxels marked as "white balls" and b is the set of all voxels marked as "black balls" from the blocking scheme, see Figure 2b.

By utilizing self-normalizing importance sampling [23, Chapter 3] it is possible to approximate the expectation by

$$\tilde{\mathbb{E}}_{\mathbf{Z}}[g(Z_i, \mathbf{x}_i, \mathbf{Z}_{-i})|\mathbf{X} = \mathbf{x}] \approx \sum_k^K \frac{\sum_{j=1}^J g(k, \mathbf{x}_i, \hat{\mathbf{z}}_{-i}^{(j)}) c(k, \hat{\mathbf{z}}_{-i}^{(j)}) \mathbb{P}(Z_i = k | \mathbf{X} = \mathbf{x}, \mathbf{Z}_{-i} = \hat{\mathbf{z}}_{-i}^{(j)})}{\sum_{j=1}^J c(k, \hat{\mathbf{z}}_{-i}^{(j)})}$$

However, approximating $\tilde{\mathbb{E}}[\dots]$ satisfactory using self normalizing importance sampling is more computationally demanding and the resulting parameter estimates are in practice very close to the ones obtained by just approximating $\tilde{\mathbb{E}}_{\mathbf{Z}}[\dots] \approx \mathbb{E}_{\mathbf{Z}}[\dots]$. Therefore, we use this approximation in favor of the importance sampling described above, as has been done before in similar problems [7, 24].

The computational complexity of approximating $\mathbb{E}_{\mathbf{Z}}[g(Z_i, x_i, \mathbf{Z}_{-i})|\mathbf{x}]$ for a general function g is hence of order $\mathcal{O}(JNK^2)$, where J is the number of Monte Carlo iterations, K is the number of classes and N is the number of voxels in the image. The N and one of

the K factors are attributed to a Gibbs sampling stage. The J and the other K factor are attributed to the sum in the Rao-Blackwellisation of the Monte Carlo integration. Note that approximating the posterior probabilities of the latent field only has a complexity of $\mathcal{O}(JNK)$ since the sum over the classes in (6) reduces to the term involving $k = z_i$ due to the indicator function.

For a valid MCMC simulation J needs to be large enough so that the burn in phase of the Gibbs simulation can be omitted. However, the EM gradient algorithm is an iterative procedure and in each iteration one needs new evaluations of $\pi(z_i|\mathbf{x})$. In order to reduce the computations one can make use of the iterative scheme of the EM gradient algorithm. J can still be chosen small since each gradient iteration feed off the MCMC sampling of the former one. The idea behind this is that the spatial field characterized by the spatial parameters (α_k, β) will be reasonably similar between two consecutive gradient iterations. One can then use the field generated in the former gradient iteration as an initial value for the new MCMC simulation. Through this trick the need for a burn in period is basically eliminated. In our implementation we used $J = 10$ for the parameter estimation phase.

B EM Gradient conditional line search

Performing a conditional line search in the EM gradient algorithm corresponds to numerically finding a value of δ that maximizes

$$Q\left(\Theta^{(j)} + \delta H^{-1}\left(\log \tilde{L}(\Theta^{(j)}; \mathbf{x})\right) \nabla \left(\log \tilde{L}(\Theta^{(j)}; \mathbf{x})\right) \middle| \mathbf{X} = \mathbf{x}; \Theta^{(j)}\right),$$

where

$$\begin{aligned} Q(\Theta|\Theta^{(j)}) &= \sum_{i=1}^N \sum_{k=1}^K \log f(x_i|Z_i = k; \Theta) \mathbb{P}(Z_i = k|\mathbf{X} = \mathbf{x}; \Theta^{(j)}) \\ &+ \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}_{\mathbf{Z}_{-i}} [\log \pi(Z_i = k|\mathbf{Z}_{-i}; \Theta)|\mathbf{X} = \mathbf{x}, \Theta^{(j)}] \mathbb{P}(Z_i = k|\mathbf{X} = \mathbf{x}; \Theta^{(j)}). \end{aligned}$$

A line search requires evaluations of the function Q for several values of δ . Fortunately, both $\log f(x_i|Z_i = k; \Theta)$ and $\mathbb{E}_{\mathbf{Z}_{-i}} [\log \mathbb{P}(Z_i = k|\mathbf{Z}_{-i}; \Theta)|\mathbf{X} = \mathbf{x}, \Theta^{(j)}]$ can be calculated explicitly within a feasible computational cost. This since the first term does only depends on the current voxel i and the second term does only need to be summed over all possible states of Z_j for the nearest neighbors to voxel i since the Z -field is Markov. This is feasible since both the number of classes and the number of neighbors are typically small. Also, the probabilities $\mathbb{P}(Z_i = k|\mathbf{X} = \mathbf{x}; \Theta^{(j)})$, are already approximated with the Monte Carlo simulation in the gradient step and can be reused without further computations, hence the line search is a tractable method to ensure convergence.

Remember the equation

$$\begin{aligned} \nabla \log \tilde{L}(\Theta; \mathbf{x}) &= \sum_{i=1}^N \mathbb{E}_{\mathbf{Z}} [\nabla \log f(\mathbf{x}_i|Z_i; \Theta)|\mathbf{X} = \mathbf{x}; \Theta] \\ &+ \sum_{i=1}^N \mathbb{E}_{\mathbf{Z}} [\nabla \log \mathbb{P}(Z_i = z_i|\mathbf{Z}_{-i} = \mathbf{z}_{-i}; \Theta)|\mathbf{X} = \mathbf{x}; \Theta] \end{aligned}$$

from Section 3.1. The two terms inside the expectation depends on two mutually exclusive sets of parameters. The first term is associated with the parameters of the mixture distributions $(\mu_k, \gamma_k, \tau_k, Q_k)$, and the second term is associated with the parameters of the MRF (β and α). Conditioned on \mathbf{z} , the gradient with regard to the MRF parameters is not dependent on the mixture distributions parameters and vice versa. Hence, the approximate M-step of the EM gradient method can be separated into two separate Newton steps, one for the MRF parameters and one for the mixture parameters. It is therefore reasonable to choose the step lengths separately for the two steps. This can be beneficial since one set of parameters might need a smaller step size in some regions of the parameter space while constraining the other parameter set to the same small step size might inhibit convergence speed.

We take advantage of this separation by performing a line search for the parameters of the mixture distributions, which often has shown to need a smaller step size than the one proposed by the approximate Newtons method. However, we simply use a fixed step length for the MRF parameters since the proposed step lengths of the approximate Newtons method seem to be satisfactory. Also the amount of indexing needed to implement the line search in this case significantly increases the execution time of the parameter estimation.

C Derivations of properties of the NIG distribution

In order to prove Propositions 2.1 and 2.2, let us recall some properties of the NIG distribution. If \mathbf{X} is NIG distributed, we have that $\mathbf{X}|V \sim \mathbf{N}(\mu - \gamma v, vQ^{-1})$, where $V \sim \text{IG}(\tau, \sqrt{\tau/2})$. The density of \mathbf{X} can thus be derived by computing

$$f(\mathbf{x}) = \int f(\mathbf{x}, v) dv = \int f(\mathbf{x}|v) f(v) dv \quad (7)$$

where $f(\mathbf{x}|v)$ is the density of $\mathbf{X}|V$ and $f(v)$ is the density of the inverse Gaussian distribution, which is shown in Appendix D.1. It is easy to evaluate the integral if one identify the factors including v as a probability density function of a GIG distribution with parameters $\nu = -\frac{d+1}{2}$, $a = \gamma^T Q \gamma + \frac{\tau}{\delta^2}$, $b = (\mathbf{x} - \mu)^T Q (\mathbf{x} - \mu) + \tau$ and without correct normalizing constant.

C.1 Proof of Proposition 2.1

We have that

$$\begin{aligned} f(\mathbf{x}^B) &= \int f(\mathbf{x}^B|v) f(v) dv = \int \left(\int f(\mathbf{x}|v) d\mathbf{x}^A \right) f(v) dv \\ f(\mathbf{x}^B|v) &= \frac{\sqrt{|\hat{Q}|}}{(2\pi v)^{(d-d_m)/2}} e^{-\frac{1}{2v}(\mathbf{x}^B - \mu^B - \gamma^B v)^T \hat{Q} (\mathbf{x}^B - \mu^B - \gamma^B v)}. \end{aligned}$$

The resulting density is now acquired by recognizing the integral as the integral over a GIG distribution without a proper normalization analogous to how the density for the joint distribution was acquired by the integral (7). This gives

$$f(\mathbf{x}^B) = \frac{\sqrt{\tau |\hat{Q}|}}{(2\pi)^{(d-d_m+1)/2}} e^{((\mathbf{x}^B - \boldsymbol{\mu}^B)^T \hat{Q} \boldsymbol{\gamma}^B + \frac{\tau}{\delta})} 2K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}}) \left(\frac{\hat{b}}{\hat{a}}\right)^{\frac{\hat{\nu}}{2}},$$

where the hat denotes parameters of the marginal distribution and relates to the original parameters as follows: $\hat{a} = \boldsymbol{\gamma}^B \hat{Q} \boldsymbol{\gamma}^B + \frac{\tau}{\delta^2}$, $\hat{b} = (\mathbf{x}^B - \boldsymbol{\mu}^B)^T \hat{Q} (\mathbf{x}^B - \boldsymbol{\mu}^B) + \tau$, $\hat{\nu} = -\frac{d-|A|+1}{2}$, and $\hat{Q} = (\Sigma^{BB})^{-1}$. We can identify this as the NIG distribution given in the proposition. \square

C.2 Proof of proposition 2.2

We have that

$$f(\mathbf{x}^A | \mathbf{x}^B) = \int f(\mathbf{x}^A | \mathbf{x}^B, v) f(v | \mathbf{x}^B) dv$$

where $f(v | \mathbf{x}^B) \propto f(\mathbf{x}^B | v) f(v)$ and

$$\mathbf{x}^A | \mathbf{x}^B, v \sim \mathcal{N} \left(\boldsymbol{\mu}^A + \boldsymbol{\gamma}^A v - (Q^{AA})^{-1} Q^{AB} (\mathbf{x}^B - \boldsymbol{\mu}^B - \boldsymbol{\gamma}^B v), v (Q^{BB})^{-1} \right)$$

From Section C.1 we know that $f(\mathbf{x}^B | v) f(v)$ corresponds to the density function of a GIG distribution without normalization with parameters $(\hat{\nu}, \hat{a}, \hat{b})$. Hence $v | \mathbf{x}^B \sim \text{GIG}(\hat{\nu}, \hat{a}, \hat{b})$ and since $\mathbf{X}^A | \mathbf{X}^B, v$ is Gaussian distributed with mean $\boldsymbol{\mu}^A + \boldsymbol{\gamma}^A v - (Q^{AA})^{-1} Q^{AB} (\mathbf{x}^B - \boldsymbol{\mu}^B - \boldsymbol{\gamma}^B v) := \mathbf{X}^A - \mathbf{L} - \mathbf{O}v$ and precision matrix $\frac{1}{v} Q^{AA}$ we get:

$$\begin{aligned} f(\mathbf{x}^A | \mathbf{x}^B) &= \int \frac{\sqrt{|Q^{AA}|}}{(2\pi v)^{|A|/2}} \left(\frac{\hat{a}}{\hat{b}}\right)^{\hat{\nu}/2} \frac{\exp\left(-\frac{1}{2v} (\mathbf{L} + \mathbf{O}v)^T Q^{AA} (\mathbf{L} + \mathbf{O}v) - \frac{\hat{a}v}{2} - \frac{\hat{b}}{2v}\right)}{2K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} v^{\hat{\nu}-1} dv \\ &\propto \int v^{\hat{\nu}-1-|A|/2} \exp\left(-\frac{1}{2v} (\mathbf{L} + \mathbf{O}v)^T Q^{AA} (\mathbf{L} + \mathbf{O}v) - \frac{\hat{a}v}{2} - \frac{\hat{b}}{2v}\right) dv \\ &\propto e^{-\mathbf{L}^T Q^{AA} \mathbf{O}} \int v^{\hat{\nu}-1-|A|/2} \exp\left(-\frac{v}{2} (\mathbf{O}^T Q^{AA} \mathbf{O} + \hat{a}) - \frac{1}{2v} (\mathbf{L}^T Q^{AA} \mathbf{L} + \hat{b})\right) dv. \end{aligned}$$

The integral can be identified as a integral over a GIG distribution without proper normalization. This give us the stated conditional density function. Moreover,

$$\mathbf{X}^A | \mathbf{X}^B \sim \text{GH}(\tilde{\boldsymbol{\mu}}, Q^{AA}, \tilde{\boldsymbol{\gamma}}, \hat{\nu}, \hat{a}, \hat{b}),$$

with the definition of a generalized hyperbolic distribution as in D.3. Here, $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}^A - (Q^{AA})^{-1} Q^{AB} (\mathbf{x}^B - \boldsymbol{\mu}^B)$ and $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^A + (Q^{AA})^{-1} Q^{AB} \boldsymbol{\gamma}^B$.

From this the conditional expectation can be derived as

$$\begin{aligned}
\mathbb{E}[\mathbf{X}^A | \mathbf{x}^B] &= \int \mathbf{x}^A f(\mathbf{x}^A | \mathbf{x}^B) d\mathbf{x}^A = \int \mathbf{x}^A \int f(\mathbf{x}^A | \mathbf{x}^B, v) f(v | \mathbf{x}^B) dv d\mathbf{x}^A \\
&= \int \mathbb{E}[\mathbf{X}^A | \mathbf{x}^B, v] f(v | \mathbf{x}^B) dv \\
&= \int \left[\boldsymbol{\mu}^A + \boldsymbol{\gamma}^A v - (Q^{AA})^{-1} Q^{AB} ((\mathbf{x}^B - \boldsymbol{\mu}^B - \boldsymbol{\gamma}^B v)) \right] f(v | \mathbf{x}^B) dv \\
&= \left[\boldsymbol{\mu}^A - (Q^{AA})^{-1} Q^{AB} (\mathbf{x}^B - \boldsymbol{\mu}^B) \right] + \left[\boldsymbol{\gamma}^A + (Q^{AA})^{-1} Q^{AB} \boldsymbol{\gamma}^B \right] \mathbb{E}[v | \mathbf{x}^B] \\
&= \left[\boldsymbol{\mu}^A - (Q^{AA})^{-1} Q^{AB} (\mathbf{x}^B - \boldsymbol{\mu}^B) \right] \\
&\quad + \left[\boldsymbol{\gamma}^A + (Q^{AA})^{-1} Q^{AB} \boldsymbol{\gamma}^B \right] \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\gamma}} \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})},
\end{aligned}$$

where second to last step used the expression for the expected value of a GIG variable, given in Appendix D.1. Similarly the conditional covariance can be derived by

$$\begin{aligned}
\mathbb{E}[(\mathbf{X}^A)^2 | \mathbf{x}^B] &= \int [\text{Cov}(\mathbf{X}^A | \mathbf{x}^B, v) + \mathbb{E}[\mathbf{X}^A | \mathbf{x}^B, v] \mathbb{E}[\mathbf{X}^A | \mathbf{x}^B, v]^T] f(v | \mathbf{x}^B) dv \\
&= \int \left[v (Q^{AA})^{-1} + (\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\gamma}} v)(\tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\gamma}} v)^T \right] f(v | \mathbf{x}^B) dv \\
&= \int \left[\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^T + (\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\gamma}}^T + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\mu}}^T + (Q^{AA})^{-1}) v + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\gamma}}^T v^2 \right] f(v | \mathbf{x}^B) dv \\
&= \tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^T + (\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\gamma}}^T + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\mu}}^T + (Q^{AA})^{-1}) \mathbb{E}[V | \mathbf{x}^B] + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\gamma}}^T \mathbb{E}[V^2 | \mathbf{x}^B] \\
&= \tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^T + (\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\gamma}}^T + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\mu}}^T + (Q^{AA})^{-1}) \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\gamma}}^T \frac{\hat{b}}{\hat{a}} \frac{K_{\hat{\nu}+2}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})}
\end{aligned}$$

where the expression for $\mathbb{E}[V^2 | \mathbf{x}^B]$ is taken from Appendix D.1.

$$\begin{aligned}
\mathbb{C}(\mathbf{X}^A | \mathbf{x}^B) &= \mathbb{E}[(\mathbf{X}^A)^2 | \mathbf{x}^B] - \mathbb{E}[\mathbf{X}^A | \mathbf{x}^B] \mathbb{E}[\mathbf{X}^A | \mathbf{x}^B]^T \\
&= \left[\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^T + (\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\gamma}}^T + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\mu}}^T + (Q^{AA})^{-1}) \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\gamma}}^T \frac{\hat{b}}{\hat{a}} \frac{K_{\hat{\nu}+2}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} \right] \\
&\quad - \left[\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\mu}}^T + (\tilde{\boldsymbol{\mu}} \tilde{\boldsymbol{\gamma}}^T + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\mu}}^T) \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\gamma}}^T \left(\frac{\hat{b}}{\hat{a}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} \right)^2 \right] \\
&= (Q^{AA})^{-1} \sqrt{\frac{\hat{b}}{\hat{a}}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} + \tilde{\boldsymbol{\gamma}} \tilde{\boldsymbol{\gamma}}^T \left[\frac{\hat{b}}{\hat{a}} \frac{K_{\hat{\nu}+2}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} - \left(\frac{\hat{b}}{\hat{a}} \frac{K_{\hat{\nu}+1}(\sqrt{\hat{a}\hat{b}})}{K_{\hat{\nu}}(\sqrt{\hat{a}\hat{b}})} \right)^2 \right].
\end{aligned}$$

The desired result is now obtained by using the equality

$$\frac{\nu}{z} K_{\nu}(z) - K_{\nu+1}(z) = -\frac{\nu}{z} K_{\nu}(z) - K_{\nu-1}(z)$$

for the modified Bessel function $K_{\hat{\nu}+2}(\sqrt{\hat{a}\hat{b}})$. \square

D Distributions

D.1 The generalized inverse Gaussian distribution

A random variable X has a GIG distribution if it has probability density function

$$f(x) = \left(\frac{a}{b}\right)^{\nu/2} \frac{x^{\nu-1}}{2K_\nu(\sqrt{ab})} \exp\left(-\frac{ax}{2} - \frac{b}{2x}\right).$$

The following expectations holds true for the GIG distribution,

$$\begin{aligned}\mathbb{E}[X] &= \sqrt{\frac{b}{a}} \frac{K_{\nu+1}(\sqrt{ba})}{K_\nu(\sqrt{ba})} \\ \mathbb{E}[X^2] &= \frac{b}{a} \frac{K_{\nu+2}(\sqrt{ab})}{K_\nu(\sqrt{ab})}\end{aligned}$$

D.2 The inverse Gaussian distribution

The IG is a special case of a GIG distribution with parameters $\nu = -\frac{1}{2}$, $a = \frac{\tau}{\delta^2}$, $b = \tau$, and thus has density

$$f(x) = \frac{\sqrt{\tau}}{\sqrt{2\pi x^3}} \exp\left(-\frac{\tau(x - \delta)^2}{2\delta^2 x}\right).$$

D.3 The generalized hyperbolic distribution

The generalized hyperbolic distribution (GH) is defined similarly to the NIG distribution [28], but instead of letting the latent variance variable be IG distributed it is instead GIG distributed.

$$\mathbf{X} \sim \text{GH}(\boldsymbol{\mu}, Q, \gamma, \nu, a, b) \text{ if } \begin{cases} \mathbf{X} = \boldsymbol{\mu} + \gamma V + \sqrt{V} Q^{-\frac{1}{2}} \mathbf{Z} \\ V \sim \text{GIG}(\nu, a, b) \\ \mathbf{Z} \sim \text{N}(\boldsymbol{\mu} = \mathbf{0}, \mathbf{I}) \end{cases}$$

E Continuous Ranked Probability Score

The Continuous Ranked Probability Score (CRPS) is a scoring rule that assesses how well a continuous probability distribution explains observed data. It is a proper scoring rule, i.e. the expected score is largest for the distribution from which the data actually was sampled from. It is defined as

$$\text{CRPS}(F, x) = - \int_{-\infty}^{\infty} (F(y) - \mathbb{I}(y \geq x))^2 dy,$$

where F is the cumulative distribution function of a distribution and x is some observation. Often CRPS is used negatively oriented ($\text{CRPS}^* = -\text{CRPS}$) since then it is positive and a small value close to zero corresponds to a good fit.

Gneiting and Raftery [12] showed that the CRPS can be expressed as

$$CRPS(F, x) = \frac{1}{2} \mathbb{E} [|Y - Y'|] - \mathbb{E} [|Y - x|],$$

where Y and Y' are i.i.d. with a distribution that corresponds to F .

If F corresponds to a normal distribution, then both $|Y - Y'|$ and $|Y - x|$ are distributed as a folded normal distribution. If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then one has

$$\mathbb{E}[|Y|] = 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right). \quad (8)$$

In this work we need to compute the CRPS for mixture models with the class membership field Z and the conditional distribution $X_i^A | \mathbf{X}^B = \mathbf{x}^B$. Hence

$$\begin{aligned} CRPS(F, x_i^A) &= \mathbb{E} \left[\frac{1}{2} \mathbb{E} [|X_i^A - X_i^{A'}| | Z_i] - \mathbb{E} [|X_i^A - x_i^A| | Z_i] \middle| \mathbf{X}^B = \mathbf{x}^B \right] \\ &= \sum_{k=1}^K \left(\frac{1}{2} \mathbb{E} [|X_i^A - X_i^{A'}| | Z_i] - \mathbb{E} [|X_i^A - x_i^A| | Z_i] \right) \mathbb{P}(Z_i = k | \mathbf{X}^B = \mathbf{x}^B), \end{aligned}$$

The posterior class probabilities, $\mathbb{P}(Z_i = k | \mathbf{X}^B = \mathbf{x}^B)$, are already approximated by MCMC simulation during the s-CT prediction, see Appendix A. For the GMM and GMMS models the CRPS can therefore be computed explicitly given those probabilities and equation (8) since conditioned on Z_i , $X_i^A | \mathbf{X}^B$ is normally distributed.

For the NIG mixtures we have no explicit expression of the CRPS and instead we need to compute them by Monte Carlo simulations. Just as in Bolin and Wallin [6] the variance of the MC simulation can be significantly decreased by realizing that conditioned on the class and the variance variable, V_i , the NIG distribution is also normally distributed. One therefore need to Monte Carlo simulate the variance variable, but not the entire NIG variable, since the CRPS value of a normally distributed variable could be acquired analytically. Thus,

$$\begin{aligned} CRPS(F, x_i^A) &= \frac{1}{2n} \sum_{k=1}^K \sum_{j=1}^n \left(\mathbb{E} [|X_i^A - X_i^{A'}| | V_i = v_j, V_i' = v_j', Z_i = k] \right. \\ &\quad \left. - \mathbb{E} [|X_i^A - x_i^A| | V_i = v_j, Z_i = k] \middle| Z_i = k \right) \cdot \mathbb{P}(Z_i = k | \mathbf{X}^B = \mathbf{x}^B), \end{aligned}$$

where v_j and v_j' are sampled from the current $V_i | Z_i$.